STATEMENT OF INTEREST

Community for Data Integration Funding FY2014

SECTION 1. PROJECT ADMINISTRATIVE INFORMATION

- **This proposal is in response to CDI SSF Category 4**: Community Innovation
- **Project Title**: Characterization of earthquake damage and effects using social media data
- **USGS Cost Center requesting funding**: Geological Hazards Science Center, Golden CO.
- **USGS principal investigator**: Michelle Guy; Computer Scientist; P.O. Box 25046, MS 966 DFC, Denver CO 80225-0046; office 303-273-8650; fax 303-273-8450; mguy@usgs.gov
- **Additional USGS principal investigator and collaborators**: Paul Earle; Geophysicist and NEIC Operations Lead; P.O. Box 25046, MS 966 DFC, Denver CO 80225-0046; office 303-273-8417; Scott Horvath; Bureau Social Media Coordinator; 12201 Sunrise Valley Drive, MS 119, Reston, VA 20192; office 703.648.4011; shorvath@usgs.gov
- **Non USGS collaborator**: Doug Bausch; Senior Physical Scientist, Mitigation Division, FEMA Region VIII; Denver Federal Center, Bldg. 710a, Box 25267, Denver, CO 80225-0267; office 303-235-4859; Douglas.Bausch@dhs.gov;
- **Project description**: Add characterization of earthquake damage and effects, using multiple social-media sources, to an existing application resulting in rapidly generated geo-referenced impact reports for scientists and an innovative and unique dataset of raw and processed qualitative social-media data for integration with traditional seismic data.

SECTION 2. PROJECT SUMMARY

We propose an innovative use of social media data to rapidly characterize damage and effects following significant earthquakes. First hand witnessed earthquake effects reported via social-media technologies are significantly faster, on the order of seconds to hours, than the time it takes for first responders, reporters and scientists to travel into impacted areas. In this project we will acquire and process witnessed effects reported via Twitter, Instagram, and Flickr and then rapidly make the data and derived products available to scientists and potentially first responders as summarized geo-referenced impact reports. These data will be integrated with traditional seismic data to cost effectively supplement the USGS National Earthquake Information Center's seismically derived impact estimates. Integration of these data sources will provide important verification of the qualitative social media data which often has increased spatial coverage and detail. The project builds on an existing dataset and operational system that rapidly detects on average three widely felt earthquakes per day using Twitter posts.

The existing operational system, Tweet Earthquake Dispatch (TED), collects tweets containing the word "earthquake" and its equivalent in several languages, then archives these data and performs real-time cluster analysis to make event detections [Guy et al., 2010], with a 95% positive identification rate. In regions of sparse seismic station coverage, TED often detects earthquakes seconds to minutes before seismically confirmed data is processed and made available to USGS scientists. During the system's three years of operation, TED has amassed a dataset of over 22 million "earthquake" tweets that have yet to be rigorously analyzed. Twitter only permits searches of their tweet archive for the past seven days so the USGS archive is a unique dataset for testing data mining and real-time algorithms.

We propose expanding TED beyond its current real-time detections to include rapid characterization of earthquake damage and effects by 1) expanding our social-media dataset to include tweets following an event that contain other damage related words add data from other social media sources such as Instagram and Flickr, 2) using our existing dataset to develop algorithms to automatically cull and condense the enormous number of tweets produced after a significant earthquake into a summary report and 3) build a real-time prototype system to distribute the earthquake effect reports to scientists and emergency responders for evaluation and possible implementation in a robust operational system. The data collected for this study is highly cost effective as there are no field instruments to purchase, deploy or maintain and the data stream is freely available. The data will be shared to the extent possible given restrictions imposed by the social media sources. Other groups such as FEMA and international earthquake monitoring centers are actively interested in the current and proposed datasets.

Products expected from this work include 1) an extension to the existing database schema creating a richer dataset, 2) extension of the existing system for additional data analysis and automated report generation 3) a mobile friendly distribution of impact reports and 4) integration of data and derived products with a traditional seismic catalog for long term storage, reference and result validation.

The project is multidisciplinary by nature and the PIs include experts in computer science (Guy, USGS), seismology (Earle, USGS), social media (Horvath, USGS), and emergency response (Bausch, FEMA). The proposed project scope and expected products are ambitious but attainable given the foundation of the current operational system. To eliminate costly and time consuming start-up efforts the software implementation work will be done by the USGS Fort Collins development group, who built the original system, and will leverage existing hardware and GHSC IT support staff.

Reference: M Guy, P Earle, C Ostrum, K Gruchalla, S Horvath, 2010, Integration and dissemination of citizen reported and seismically derived earthquake information via social network technologies. Advances in Intelligent Data Analysis IX, 42-53

SECTION 3. ESTIMATED BUDGET

Budget Category	Federal Funding "Requested"	Matching Funds "Proposed"
1. SALARIES (including Benefits):		
Personnel Total: Analysis, Design and IT Support	--	$18,000
Contract Personnel Total: Software Implementation and Testing	$48,800	--
Total Salaries:	$48,800	$18,000
2. TRAVEL EXPENSES:		
Travel Total: One trip for Horvath from D.C .to Golden	$1,200	--
Total Travel Expenses:	$1,200	--
3. OTHER DIRECT COSTS: (itemize)		
Equipment (development and production hardware and infrastructure):	--	$5,000
Total Other Direct Costs:	--	$5,000
GRAND TOTAL	$50,000	$23,000